RESEARCH METHODOLOGY

Reasons or excuses for avoiding meta-analysis in forest plots

Heterogeneous data are a common problem in meta-analysis. **John Ioannidis**, **Nikolaos Patsopoulos**, and **Hannah Rothstein** show that final synthesis is possible and desirable in most cases

Some systematic reviews simply assemble the eligible studies without performing meta-analysis. This may be a legitimate choice. However, an interesting situation arises when reviews present forest plots (quantitative effects and uncertainty per study) but do not calculate a summary estimate (the diamond at the bottom). These reviews imply that it is important to visualise the quantitative data but final synthesis is inappropriate. For example, a review of sexual abstinence programmes for HIV prevention claimed that owing to "data unavailability, lack of intention-to-treat analyses, and heterogeneity in programme and trial designs… a statistical meta-analysis would be inappropriate."[1] As we discuss, options almost always exist for quantitative synthesis and sometimes they may offer useful insights. Reviewers and clinicians should be aware of these options, reflect carefully on their use, and understand their limitations.

Why meta-analysis is avoided

Of the 1739 systematic reviews that included at least one forest plot with at least two studies in issue 4 of the *Cochrane Database of Systematic Reviews* (2005), 135 reviews (8%) had 559 forest plots with no summary estimate.

The reasons provided for avoiding quantitative synthesis typically revolved around heterogeneity (table 1). The included studies were thought to be too different, either statistically or in clinical (including methodological) terms. Differences in interventions, metrics, outcomes, designs, participants, and settings were implied.

How large is too large heterogeneity?

This question of lumping versus splitting is difficult to answer objectively for clinical heterogeneity. Logic models based on the PICO (population-intervention–comparator-outcomes) framework may help to deal with the challenges of deciding what to include and what not. Still, different reviewers, readers, and clinicians may disagree on the (dis)similarity of interventions, outcomes, designs, participant characteristics, and settings.

No widely accepted quantitative measure exists to grade clinical heterogeneity. Nevertheless, it may be better to examine clinical differences in a meta-analysis rather than use them as a reason for not conducting one. For example, a review identified 40 trials of diverse interventions to prevent falls in elderly people.[2] Despite large diversity in the trials, the authors did a meta-analysis and also examined the effectiveness of different interventions. The analysis suggested that evidence was stronger for multifactorial risk assessment and management programmes and exercise and more inconclusive for environmental modifications and education.

Statistical heterogeneity can be measured—for example, by calculating I^2 and its uncertainty.[3-5] I^2, the proportion of variation between studies not due to chance, takes values from 0 to 100%. In the 22 forest plots including four or more studies that avoided synthesis because of heterogeneity, I^2 ranged between 35% and 98% with a median of 71% (figure). Yet, 86 of the 1011 forest plots where reviewers had no hesitation in performing meta-analysis had I^2 exceeding 71%.[5] The lower 95% confidence limit of I^2 was <25% in 11 of the 22 non-summarised forest plots—that is, for half of them we cannot exclude that statistical heterogeneity is limited. Therefore, even for statistical heterogeneity, there is substantial variability in what different reviewers consider too much. Statistical heterogeneity alone is a weak and inconsistently used argument for avoiding quantitative synthesis.

Potential methods for use in heterogeneity

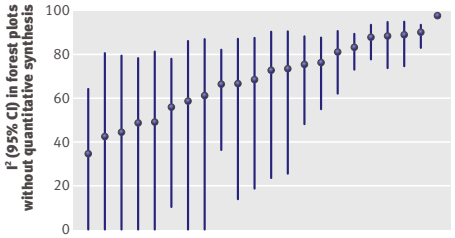
Table 2 provides methodological approaches to quantitative synthesis of data that some researchers may deem unsuitable for meta-analysis. It is unknown whether researchers preparing systematic reviews were aware of these methods but thought that they were inapplicable; were aware of their existence

Table 1 | Reasons for not showing summary estimates in forest plots from systematic reviews in Cochrane database 2005 issue 4

Reason	No (%) of systematic reviews (n=135)*
Statistical heterogeneity too high	32 (24)
Different interventions compared	41 (30)
Different metrics or outcomes evaluated	26 (19)
Different metric of same outcome	7
Different outcome	20
Different study designs	21 (16)
Non-randomised studies	3
Other design issues	18
Different study participants, settings	21 (16)
Data with many counts per participant	5 (4)
Data too limited	11 (8)
Clinical heterogeneity (not otherwise specified)	5 (4)
Synthesis considered inappropriate (not specified why)	3 (2)
Non-normality of data	1 (1)
No reason given	10 (7)
Artefact†	3 (2)
Quantitative synthesis given in text	7 (5)

*For another 7 reviews forest plots listed the names of two or more studies but data were available for only one or no study; 4 reviews listed as withdrawn are not included. Several reviews gave more than one reason without clarifying which was the most important. In these cases, all reasons are counted. However, when only one reason specifically was used to decide whether data synthesis would be appropriate, only this reason is counted, regardless of whether several other reasons were then secondarily probed.

†Meta-analyses of individual level data that used the RevMan forest plot function to show subgroup summary estimates.



I² point estimates and 95% CI for forest plots with at least 4 studies and no quantitative synthesis because of perceived high statistical heterogeneity

but lacked the necessary experience and software; or were unaware of their existence. Detailed discussion of methods is beyond our scope here, but we present the principal options and caveats and provide references for interested readers. Some methods are experimental and extra caution is needed.

Models that can accommodate statistical heterogeneity between studies include traditional random effects (models that assume that different studies have different true treatment effects),[6] meta-regressions (regressions that examine whether the treatment effect is related to one or more characteristics of the studies or patients),[7] and bayesian methods (methods that combine various prior assumptions with the observed data).[8] Random effects do not explain the heterogeneity: they distort estimates when large versus smaller studies differ in results and smaller studies are more biased, and they can be unstable with limited evidence[9]; meta-regressions may suffer from post hoc selection of variables, the ecological fallacy, and poor performance with few studies[10]; and bayesian results may depend on prior specifications.[8] Meta-analysis of data at the individual level may permit fuller exploration of heterogeneity, but these data are usually unavailable.[11]

The availability of multiple interventions for the same condition and indication is increasingly common. Different regimens may be merged in common groups, but differences in treatment effects of merged regimens may remain unrecognised. Multiple treatments meta-analysis could be used to examine all the different treatments used for a given condition. For example, 242 chemotherapy trials are available covering 137 different regimens for advanced colorectal cancer.[12] The number of possible comparisons is prohibitive. A meta-analysis grouped these regimens into 12 treatment types and then performed a network analysis that evaluated their relative effectiveness. Instead of taking one comparison at a time, the network considered concomitantly all the data from all relevant comparisons. Networks integrate information from both direct and indirect comparisons of different treatments.[13-15] Main caveats include possible inconsistency in results between direct and indirect comparisons and the still limited experience on networks.[13-16]

Clinical trials on the same topic also commonly use many different outcomes. Meta-analysis of one outcome at a time offers a fragmented picture. Some outcomes simply differ in their measures—for example, global clinical improvement measured on a continuous scale or as a binary end point (yes/no). Continuous scales can be converted into binary ones and standardised metrics (popular in the social sciences)[17] can accommodate different outcomes that measure the same construct (such as various psychometric scales). However, for medical applications, many clinicians think that anything other than plain absolute risk is insufficiently intelligible to inform practice and policy.[18 19] Finally, some outcomes may represent truly different end points with partial correlation among themselves (for example, serum creatinine, creatinine clearance, progression to end stage renal disease, initiation of renal replacement therapy) and multivariate meta-analysis models can cater for two or more correlated outcomes.[20-22] Such models borrow strength from all the available outcomes across trials. The main caveats are specification of correlations and sparse data.

The combination of data from randomised and non-randomised studies is possible using traditional meta-analysis models. The main caveats are the spurious precision,[23] confounding, and potentially stronger selective reporting biases in observational studies.[24] However, the generalised synthesis of both randomised and non-randomised studies on the same topic may offer complementary information.[25-27] Other designs that require special care in meta-analysis include cluster[28] and crossover trials.[29]

Appropriate methods also exist for synthesising data when each participant may count many times in the calculations (multiple periods at risk or multiple follow-up data).[17 30]

The authors of several systematic reviews state only that "data synthesis is inappropriate"

Table 2 | Methodological approaches to consider in the synthesis of heterogeneous data

Problem	Possible methodological solution	Selected key caveats
High statistical heterogeneity*	Random effects	Does not explain heterogeneity, small study effects, limited data
	Meta-regression	Choice of variables, ecological fallacy, limited data
	Bayesian meta-analysis	Prior specification
	Bayesian meta-regression	Similar to meta-regression and Bayesian meta-analysis
	Meta-analysis of individual level data	Unavailable individual level information
Different interventions compared	Merge interventions in same class	Unrecognised heterogeneity
	Network meta-analysis	Inconsistency in direct versus indirect comparisons
Different metrics of same outcome	Conversion formulas	Difficulties in clinical interpretation
Different outcomes, same construct	Standardised effects	Difficulties in clinical interpretation
Different outcomes	Meta-analysis of multiple outcomes	Specification of correlations
Observational data	Generalised evidence synthesis	Spurious precision, confounding, selective reporting
Cluster randomised trials	Account for clustering correlation	Unavailable sufficient information
Crossover trials	Account for period or carry-over effect	Unavailable sufficient information
Other study design issues	Same as for high statistical heterogeneity	As for high statistical heterogeneity above
Different participants or settings	Same as for high statistical heterogeneity	As for high statistical heterogeneity above
Many counts per participant	Meta-analysis of multiple period or follow-up	Unavailable sufficient information
Limited data	Standard meta-analysis methods	Caution needed as for any meta-analysis

Popular software such as RevMan can accommodate only random effects calculations, while Comprehensive Meta-Analysis also accommodates simple meta-regressions. Bayesian models and models of multiple treatments or outcomes can be run in WinBugs. Most models can also be run in STATA or R.

*The approach used for high statistical heterogeneity may also be applicable to situations where clinical heterogeneity is considered high because of differences in interventions, metrics, outcomes, designs, participants, or settings.

or allude vaguely to "clinical heterogeneity." Specifying the reasons would improve transparency of the implicit judgments. Finally, some reviews argue that data are too limited. However, meta-analysis is feasible even with two studies. For most medical questions, only few studies exist. Limited data typically yield uncertain estimates, but the quantitative accuracy of meta-analysis may actually be a reason to avoid narrative interpretation without synthesis. Limited data may also result from asking questions that are too narrow, trying to make data too similar before inclusion in the same forest plot. Forced similarity may fragment information; it is almost unavoidable that trials will differ in at least minor ways.

To synthesise or not?

If the limitations of these methods are properly acknowledged, the use of quantitative synthesis may be preferable to qualitative interpretation of the results, or hidden quasi-quantitative analysis—for example, judging studies based on P values of single studies being above or below 0.05. Such an approach can actually lead to the wrong conclusion, especially when statistical power is low.[31] For example, if an intervention is effective but two studies are done with 40% power each, the chance of both of them getting a significant result is only 16%.

More complex "home made" qualitative rules may further compound the methodological problems. This applies not only to reviews that avoid the final synthesis but also to entirely narrative reviews without any forest plots. For example, the reviewers of interventions to promote physical activity in children and adolescents "used scores to indicate effectiveness—that is, whether there was no difference in effect between control and intervention group (0 score), a positive or negative trend (+ or –), or a significant difference (P<0.05) in favour of the intervention or control group (++ or —, respectively) . . . If at least two thirds (66.6%) of the relevant studies were reported to have significant results in the same direction then we considered the overall results to be consistent."[32] Such rules have poor performance validity.

Meta-analysis is often understood solely as a means of combining information to produce a single overall estimate of effect. However, one of its advantages is to assess, examine, and model the consistency of effects and improve understanding of moderator variables, boundary conditions, and generalisability.[8 33] Different patients and different studies are unavoidably heterogeneous. This diversity and the uncertainty associated with it should be explored whenever possible. Obtaining estimates of treatment effect (rather than simple narrative evaluations) may allow more rational decisions about the use of interventions in specific patients or settings. More sophisticated methods may also capture and model uncertainties more fully and thus may actually reach more conservative conclusions than more naive approaches. However, it is then essential that their assumptions and limitations are clearly stated and inferences drawn cautiously. Any meta-analysis method, simple or advanced, may be misleading, if we don't understand how it works.

SUMMARY POINTS

Some reviews extract numerical data and generate forest plots but avoid meta-analysis

The typical reason for not doing meta-analysis is high heterogeneity across studies

Appropriate quantitative methods exist to handle heterogeneity and may be considered if their assumptions and limitations are acknowledged

Narrative summaries may sometimes be misleading

John P A Ioannidis professor
jioannid@cc.uoi.gr

Nikolaos A Patsopoulos research fellow, Clinical Trials and Evidence Based Medicine Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine and Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina 45110, Greece

Hannah R Rothstein professor, Department of Management, Zicklin School of Business, City University of New York, New York, NY 10010, USA

Accepted: 30 March 2008

Contributors and sources: The authors have a longstanding interest in meta-analysis and sources of heterogeneity in clinical research. JPAI had the original idea for the survey. NAP and JPAI extracted the data for the survey and NAP also did the statistical heterogeneity analyses. HRR rekindled the interest in pursuing the project further and the discussion evolved with interactions between JPAI, NPA, and HRR. We thank Iain Chalmers and Alex Sutton for comments on the manuscript. JPAI wrote the manuscript and the coauthors commented on it and approved the final draft.

Competing interests: None declared.

1 Underhill K, Montgomery P, Operario D. Sexual abstinence only programmes to prevent HIV infection in high income countries: systematic review. *BMJ* 2007;335:248.
2 Chang JT, Morton SC, Rubenstein LZ, Mojica WA, Maglione M, Suttorp MJ, et al. Interventions for the prevention of falls in older adults: systematic review and meta-analysis of randomised clinical trials. *BMJ* 2004;328:680.
3 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
4 Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58.
5 Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ* 2007;335:914-6.
6 Ades AE, Lu G, Higgins JPT. The interpretation of random-effects meta-analysis in decision models. *Med Dec Making* 2005;25:646-54.
7 Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999;18:2693-708.
8 Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* 1995;14:2685-99.
9 Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998;351:123-7.
10 Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21:1559-73.
11 Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clin Trials* 2005;2:209-17.
12 Golfinopoulos V, Salanti G, Pavlidis N, Ioannidis JP. Survival and disease-progression benefits with treatment regimens for advanced colorectal cancer: a meta-analysis. *Lancet Oncol* 2007;8:898-911.
13 Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. *Health Technol Assess* 2005;9.
14 Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004;23:3105-24.
15 Salanti G, Higgins J, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res* 2008;17:279-301.
16 Ioannidis JPA. Indirect comparisons: the mesh and mess of clinical trials. *Lancet* 2006;368:1470-2.
17 Cooper HM, Hedges LV, eds. *The handbook of research synthesis*. New York: Russell Sage Foundation, 1994.
18 Schünemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. *Health Qual Life Outcomes* 2006;4:62.
19 Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle N. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. *CMAJ* 1995;152:351-7.
20 Berkey CS, Anderson JJ, Hoaglin DC. Multiple-outcome meta-analysis of clinical trials. *Stat Med* 1996;15:537-57.
21 Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Stat Med* 1998;17:2537-50.
22 Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Stat Med* 2007;26:78-97.
23 Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ* 1998;316:140-4.
24 Kavvoura FK, Liberopoulos G, Ioannidis JP. Selection in reported epidemiological risks: an empirical assessment. *PLoS Med* 2007;4:e79.
25 Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286:821-30.
26 Sutton AJ, Abrams KR, Jones DR. Generalized synthesis of evidence and the threat of dissemination bias. The example of electronic fetal heart rate monitoring (EFM). *J Clin Epidemiol* 2002;55:1013-24.
27 Shrier I, Boivin JF, Steele RJ, Platt RW, Furlan A, Kakuma R, et al. Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *Am J Epidemiol* 2007;166:1203-9.
28 Campbell MK, Elbourne DR, Altman DG, CONSORT group. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;328:702-8.
29 Curtin F, Elbourne D, Altman DG. Meta-analysis combining parallel and cross-over clinical trials. III. The issue of carry-over. *Stat Med* 2002;21:2161-73.
30 Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Stat Med* 2007;26:3681-99.
31 Pettiti D. *Meta-analysis, decision analysis and cost effectiveness analysis*. 2nd ed. New York: Oxford University Press, 1999.
32 Van Sluijs EM, McMinn AM, Griffin SJ. Effectiveness of interventions to promote physical activity in children and adolescents: systematic review of controlled trials. *BMJ* 2007;335:703.
33 Berlin JA. Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *Am J Epidemiol* 1995;142:383-7.